# Faciliter la gestion des grands jeux de données


## Une plateforme pour Étiqueter - Trouver - Partager


**CONTEXTE :**

**L'augmentation de la production de données issues de la recherche est exponentielle**

**Les enjeux évoluent de la <u>production</u> vers <u>l'exploitation</u> des données.**

**Il faut <u>adapter</u> nos pratiques pour une meilleure gestion de nos données pour en permettre une <u>meilleure utilisation.</u>**

**La solution doit prendre en compte une recherche collaborative (plusieurs participants, plusieurs lieux) et multidisciplinaire (plusieurs types de données) et pouvoir traiter de grandes quantité de données (Big Data, -omics, data déluge… ).**

**CONSTAT :**

**Pour les grands jeux de données, une organisation sous forme d'arborescence est peu adaptée**

**ex : envisagez-vous de rechercher quelque chose sur le web en parcourant une <u>arborescence</u> ➤**


**SOLUTION :**

**Une plateforme pour l'étiquetage des données permettant de s'affranchir d'une arborescence unique et figée**

**Chaque fichier de données peut recevoir autant de tags que nécessaire pour le rendre <u>compréhensible</u> et <u>réutilisable</u>**

**↳ idéalement par n'importe qui**

**ÉTIQUETAGE :**

**Les tags sont des métadonnées (données sur la donnée). Une recherche multi-critères permet de retrouver facilement les données.**

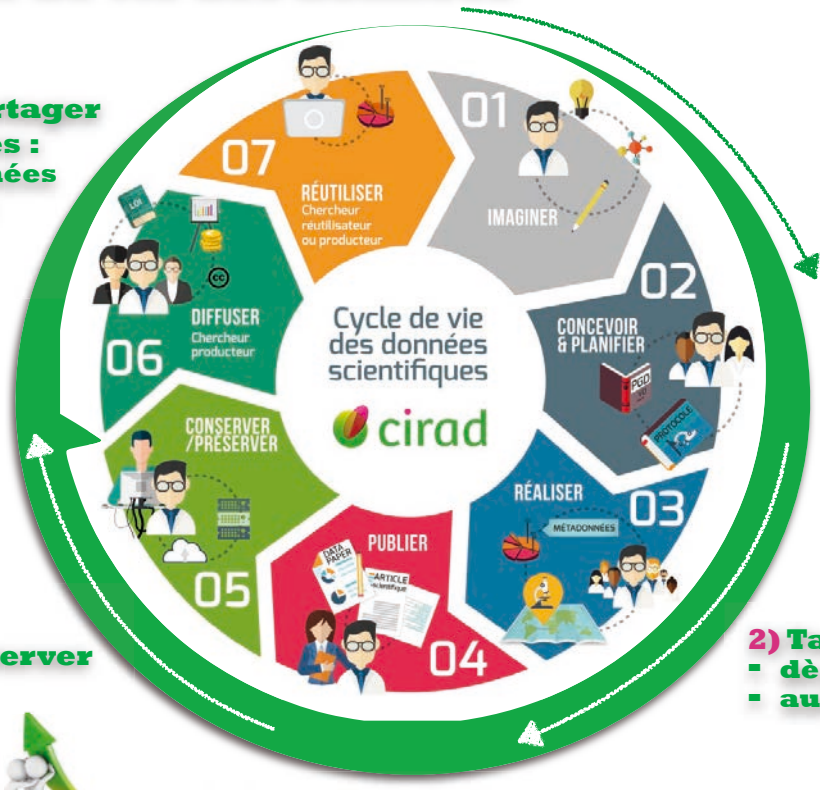
**Ex : *nom du projet, de la plante, du bailleur de fonds, du producteur de la donnée,***

***date, lieu, méthode d'obtention de la donnée,***

***type de données, format, langue, relation avec d'autres données***

**On peut aussi collecter les tags automatiquement (ex : données EXIF des photos numériques) grace à iRODS**

## Tag, métadonnée et cycle de vie des données




**5) Organiser et partager les jeux de données :**
- **Entrepôts de données**
- **Bases de données**
- **Data papers**

ORCID Connecting Research and Researchers · doi · DataCite FIND, ACCESS, AND REUSE DATA · The Dataverse Project

**1) Définir les tags :**
- **en cohérence avec la communauté scientifique**
- **avec les membres du projet**

**2) Tagger les données :**
- **dès qu'elles arrivent sur la plateforme**
- **automatiquement ou manuellement**

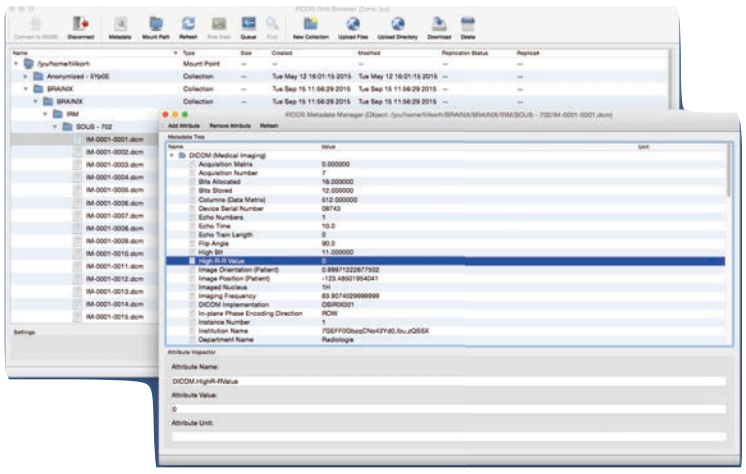
**3) Enrichir les métadonnées**

**4) Extraire et trier les données à préserver**

## Comment tagger les grands jeux de données ?

**iRODS - la solution pratique pour vos données : integrated Rule Oriented Data System - <u>irods.org</u>**
**iRODS est un « middleware » qui travaille entre vous et vos données. Il permet de définir des tags, de les associer à vos fichiers, de faire des recherches en les combinant pour partager finement vos données.**

**Ex : un fichier et ses métadonnées**



**Les 4 piliers d'iRODS**

**<u>Data Discovery</u>**
**on peut trouver les données à partir des tags**

**<u>Secure Collaboration</u>**
**on peut partager les données**

**<u>Storage Virtualization</u>**
**on peut déplacer les données sans perturbation**

**<u>Workflow Automation</u>**
**on peut surveiller la santé des données automatiquement**

**iRODS au cirad <u>800 To de stockage sécurisé</u>**
- **Pour les données de grands volumes (dès 1 To)**
- **En cours de validation sur le plateau de bioinformatique** agap

Raw data → Processed data → Information → Knowledge → Wisdom

Frédéric de Lamotte & Bertrand Pitollat UMR Agap